Data, Knowledge, & Information in Database and Knowledge-Based Systems*

Roger H. L. Chiang
Bitnet: chiang@uorgsm

Terence M. Barron
Bitnet: barron@uorgsm

Veda C. Storey
Bitnet: storey@uorgsm

William E. Simon Graduate School of Business Administration
University of Rochester
Rochester, NY 14627
Fax: 716-271-8752

March 11, 1992

Forthcoming: *The Third International Conference of the Information Resources Management Association* (May 1992)

*This research was supported by the William E. Simon Graduate School of Business Administration, University of Rochester.

Data, Knowledge, & Information in Database and Knowledge-Based Systems

Abstract

Although the terms "data", "knowledge", and "information" are frequently used when referring to information systems, there is neither a clear distinction amongst them, nor a clear way to apply them in the development of database and knowledge-based systems. The objective of this paper is to analyze the current literature on data, knowledge and information, and strive to provide a set of guidelines for distinguishing amongst these terms. These guidelines could be useful to system users and developers to make efficient information resources management and utilization. In order to do so, semiotics is employed as a framework for analyzing existing interpretations of these terms. As a result of the analysis, five prominent features are identified which are then used to discuss data, knowledge, and information.

1 INTRODUCTION

Current practices in the development of information systems suggest that the first, and most important step in the development process involves understanding the "domain" of the problem[1]. Fundamental to this understanding is the adoption of a common vocabulary that is meaningful to both the system's users and its developers. It is, therefore, remarkable that, although the terms database system, knowledge-based system, and information system are used universally, the terms, "data", "knowledge" and "information" themselves are rarely defined precisely and little agreement exists about the domain and scope encompassed by each [Wiederhold, 1986b; Bubenko and Orci, 1989]. Knowledge, for example, is often used as a synonym for either data or information. Information is usually used as the most general term and can mean either data or knowledge [Stamper, 1973]. If, however, data, knowledge, and information are different things with different properties, then one would expect different design issues to arise and different methods to be used in developing and implementing systems that must process them.

In this paper, the term "information system" refers to any system whose inputs and outputs consist of "signs"[2]. Users of an information system provide input relating to the real world

[1] This is referred to, for example, as the *requirements specification and analysis phase* in the database design process.

[2] Signs include numerical and alphabetical characters, words, sentences, messages of any length, and actions [Stamper, 1973].

and/or use the outputs as a basis from which to make decisions. The outputs of an information system can also be used as inputs to other information systems or to the system itself. Figure 1 shows a generic information system model.

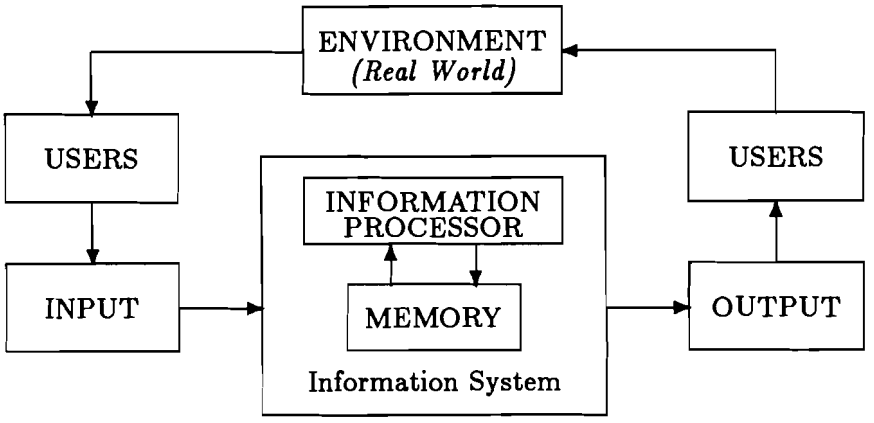

Figure 1: A generic information system model

This research strives to provide an appropriate set of characteristics for distinguishing data, knowledge, and information, which could be useful to system users and developers. In order to carry out this task, existing interpretations of data, knowledge, and information are presented and examined using semiotics as a general framework for analysis. The paper is divided into four sections. Section 2 introduces and applies the semiotics framework. Suggestions for characterizing data, knowledge, and information are then presented in Section 3. Section 4 summarizes and concludes the paper.

2 SEMIOTICS FRAMEWORK

Data, knowledge, and information (as naively understood) require signs for their representation and processing by computers. Therefore, an analysis of the properties of signs is a crucial first step in understanding the differences amongst these terms. Since *semiotics* is concerned with the properties of things in their capacity as signs [Morris, 1946], we have chosen to adopt it as a framework for our analysis. Semiotics is usually divided into three branches: *pragmatics*, *semantics*, and *syntactics* [Morris, 1946]. Pragmatics deals with the origin, uses, and effects of signs within the behaviour in which they occur and, thus, explicitly includes the users of the signs. Semantics deals with the signification of signs; that is, the relationship between the signs and the objects to which the signs refer and ignores the user of the signs. Syntactics

analyzes the relationships among signs without concern for the user or the signification of the signs [Carnap, 1942]. Figure 2 shows the scope of semiotics and illustrates the relationships of signs to users and real-world objects. Each branch of the semiotics framework is applied to existing interpretations of data, knowledge, and information below.

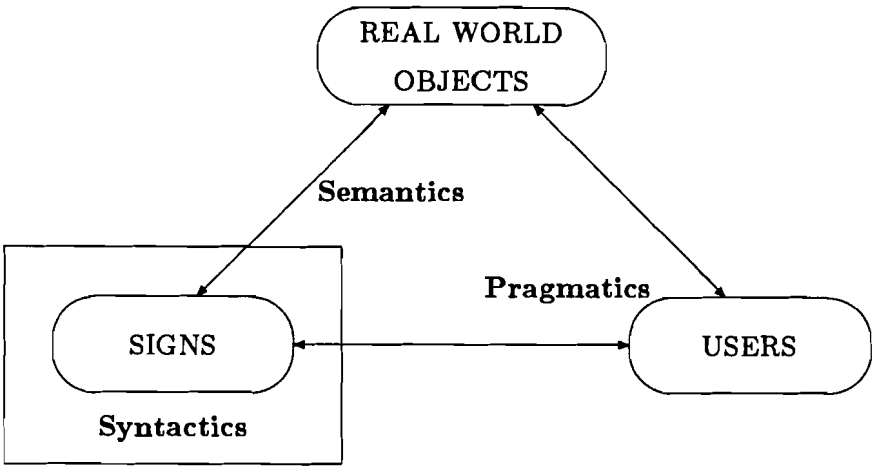

Figure 2: The Semiotics Framework

2.1 Pragmatics

Pragmatics deals with relationships between signs and behaviour [Stamper, 1973]. The way in which signs acquire their meaning is dependent on the behaviour of their users. Much of the research that focuses on the distinctions amongst data, knowledge, and information can be classified as having a pragmatic interpretation.

2.1.1 Bell

Bell [1979] distinguishes between information and knowledge based on how they should behave. Knowledge is referred to as an organized set of statements of facts or ideas which present a reasoned judgment or an experimental result. Information is defined as data processing (manipulation) in the broadest sense; that is, the result of storage, retrieval, and processing of data. This definition explicitly indicates the required operations for providing information.

2.1.2 Newell

Newell [1982] defines the following Principle of Rationality: *If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action.* This principle is clearly an extreme form of pragmatics, dealing only with the goal-seeking behaviour of an agent. In this context, *knowledge* is: "[w]hatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality. Knowledge is to be characterized entirely functionally, in terms of what it does, not structurally, in terms of physical objects with particular properties and relations." Thus, in specifying what knowledge is, it is important to understand why it is valuable and how it is used. The answers to these questions will serve as fundamental requirements for any system whose purpose is to store and process (representations of) knowledge. In addition, Newell distinguishes between: 1) knowledge itself, which is inherently abstract and exists functionally (the knowledge level); and 2) representations of knowledge which are models of knowledge (the symbol level). Furthermore, there are at least two ways in which such representations are useful: to the agent itself, as an aid in carrying out its goal-seeking behaviour, and to any observer (including the agent) who is trying to analyze the nature and extent of the knowledge possessed by the agent. Note that these two roles can have quite different implications for the nature of the representation scheme chosen. Newell also suggests that goals are a kind of knowledge having three constituents: knowledge of the desired state of affairs; knowledge that the state of affairs is desired; and knowledge of associated concerns, such as useful methods, prior attempts to attain the goals, etc.

2.1.3 Bubenko and Orci

Bubenko and Orci [1989] discuss three different views of knowledge: a database view, an Artificial Intelligence (AI) view, and an epistemic view. Their AI view is, in essence, the same as Newell's. Bubenko and Orci [1989] present a classical definition of knowledge from philosophy, which they call the epistemic view:

> C knows k iff (1) k is true, (2) C accepts k, and (3) k is evident for C

where k is a statement. This implicitly defines the knowledge of the knower, C, presumably, as the set of all k that C knows, perhaps restricted to a specific topic. Clearly, the role of the knower is an inseparable part of the definition, and, therefore, this is a pragmatics view. Although, it is rare in practical matters that truth with certainty is possible, clause (1) makes

the point that a statement, k, should be (approximately) correct with reference to its subject matter before k can be said to be known. Furthermore, truth in the sense that inference rules be truth-preserving is also a desirable property. This implies that it is desirable for the semantics of any system to be well-defined, a point that is made forcefully in Hayes [1985]. Clauses (2) and (3) of this definition imply that there is some valid chain of reasoning that leads to k as its conclusion, and that C finds this reasoning convincing.

This definition raises two important points. First, it suggests that if a human user is to accept output, k, from a system as knowledge, then it is likely that the user will need to be provided with the evidence which led the system to transmit k; that is, an explanation of the system's evidence and reasoning. Closely related to this notion of credibility is the need for truth-preserving procedures (e.g. inference rules) and for the users to understand what they are.

2.1.4 Ramamoorthy and Wan

Ramamoorthy and Wan [1989] explicitly attempt to distinguish among the three terms, data, information and knowledge. Their discussion involves pragmatics, semantics, and syntactics. With respect to data and knowledge, they state that: "Data refer to numerical information suitable for computer processing, while knowledge refers to the sum or range of what has been perceived, discovered, or learned. Knowledge can be considered data at a high level of abstraction and can be processed by a computer when it is represented as data." Their reference to abstraction is vague, but intuitively important. As a result, the notion of "abstraction" will be in more detail in Section 3. Ramamoorthy and Wan also suggest that, in general, knowledge can be considered as a compact and sometimes imprecise way of representing a body of data. This interpretation offers a "data reduction" definition of knowledge. Knowledge is the summarization of a larger body of underlying data, perhaps by statistical means. There are two major motivations for such approaches: (1) comprehensibility by humans, and (2) economizing on storage and processing. (The first of these is actually a special case of the second.) Much research has led to the conclusion that humans are very poor at making direct use of even quite small quantities of data, so that the gain from judicious summarization typically far offsets the resulting losses from inaccuracies. Thus, for example, an estimated regression equation having only a few parameters will typically be far more informative to a user than would the raw data from which those parameters were calculated.

2.1.5 Wiederhold

Wiederhold [1984; 1986b] examines the differences between data and knowledge within the context of database management systems and knowledge-based systems. He defines data as the representation of facts. Facts, in turn, are (true) statements about individual objects in some domain, provided that the truth of the statements and the gathering of the facts in the first place can be objectively verified. This suggests that the verification is performed by the senses (including inanimate sensors, for example, a temperature or pressure sensor) with minimal processing or interpretation. As a result, performing the observation and updating the database when facts change can be carried out by anyone having normal human senses. Knowledge is defined less clearly [1986b]. Wiederhold suggests that if one is looking for an expert to provide the material, then he or she is talking about knowledge. Knowledge is characterized as including abstractions and generalizations of voluminous material. Unlike data, knowledge is typically less precise and cannot be easily, objectively verified.

Wiederhold [1984] also suggests that (in terms of knowledge bases): "Knowledge is typically subjective, relates to general aspects of the data, and is significantly smaller than the data. Unlike the data, it should not vary rapidly over time, since changes require mediation by an expert." Wiederhold [1986b] defines information following Shannon and Weaver [1948] in which data is treated as a source for information, but information must convey messages that were previously unknown to the receiver. This is classified here since whether a message is information depends upon its intended receiver.

2.1.6 Summary

Table 1 summarizes pragmatic interpretations of data, knowledge, and information. It appears that data changes over time but can be verified objectively through observations of reality. Knowledge is a set of data structures and interpretative procedures that produce a certain (intelligent) behaviour in an agent. It does not vary rapidly over time. Verification requires judgment. In general, knowledge is characterized by its structures and functionalities which together produce behaviour. Information can be viewed as the symbol manipulation that is executed in order to provide something that was previously unknown to its receivers. Thus, it is relative to what the receiver knew beforehand.

Table 1: **Definitions** *(Pragmatics Approach)*

Sources	Terms	Interpretations
Bell [1979]	Knowledge	Organized set of statements of facts or ideas representing a reasoned judgment or an experimental result.
	Information	Data processing in the broadest sense.
Newell [1982]	Knowledge	Whatever can be ascribed to an agent, such that its behaviour can be computed according to the principle of rationality. Characterized entirely functionally (what it does), not how it is represented.
Bubenko and Orci [1989]	Knowledge	Collection of data structures and interpretive procedures; epistemic view.
Ramamoorthy and Wan [1989]	Data	Numerical information suitable for computer processing.
	Knowledge	Data at a higher level of abstraction.
Wiederhold [1984; 1986b]	Data	Changes rapidly over time; objectively verified.
	Knowledge	1) Should not vary rapidly over time; cannot be easily, objectively verified. 2) Precision of knowledge requires judgment.
	Information	Conveys material previously unknown to receiver.

2.2 Semantics

Semantics deals with relationships between signs and things in the "real world", especially the problems of making relationships stable and reliable enough for us to communicate with one another accurately on matters of facts and judgment [Morris, 1946].

2.2.1 Bubenko and Orci

Bubenko and Orci [1989]'s database view that attempts to distinguish data and knowledge, involves both semantics and syntactics. They consider data to be assertions about perceived states-of-affairs in the UoD (Universe of Discourse); for example, "John is married to Mary" is a perceivable or observable fact. Knowledge is considered to be at a higher level of abstraction than data; for example, properties of the UoD's states-of-affairs, general rules and relationships between individuals and facts about the UoD, how a set of facts may change, etc. Information is defined as the meaning or intentional interpretation of data. Thus, because of the emphasis on the relationship of data and knowledge to the parts of the real-world to which they refer, they are semantic notions.

2.2.2 Ramamoorthy and Wan

Ramamoorthy and Wan [1989] emphasize the relationship of data and knowledge by their meanings. Although data can be observed, the amount of data necessary to determine the truth of a piece of knowledge can be infinite. Thus, knowledge is often imprecise or uncertain.

2.2.3 Wiederhold

Wiederhold [1986a; 1986b]'s distinction between data and knowledge seems to employ a philosophical view of concepts and entities which can be described by their intension and extension. An intension is a set of criteria that something must satisfy before being included in a system. In other words, it is a template or a schema which individuals must satisfy; for example, the schema of a database. The extension is the set of things that satisfy the criteria of the intension. Under this view, data reflects properties for a particular instance, whereas knowledge conveys meanings and properties for a class of instances.

Because data reflects the state of the world at the instance level, it can involve much detail and be voluminous. When the states of instances change rapidly, data must be collected over time as well. Knowledge refers to entity types and deals with generalizations. It may be complex, but will not change frequently. The statements, "The age of student Roger is 32 years old" and "There are 52 students taking the course MIS401 during the Spring quarter of 1992" can be viewed as data, since they record facts about instances of the entity types *Student* and *Course*. The statement, "Most Ph.D. students are married", however, is considered to be knowledge since it is a statement about the entire entity type, *Ph.D. Students*. In Wiederhold's interpretation, both data and knowledge refer to the real world, but with different levels of detail. Since these interpretations consider relationships between statements and objects in the real world, they are classified under the semantics approach.

2.2.4 Summary

Table 2 summarizes interpretations of data, knowledge, and information using the semantics approach. Data seems to convey trivial meanings because they can be obtained simply through observations of reality. Knowledge contains complicated meanings obtained through a judgmental or generalization process. The meaning of a piece of knowledge is typically prone to be uncertain and imprecise, since the number of observations necessary to determine whether a piece of knowledge is true or not can be infinite in size.

Table 2: **Interpretations** *(Semantics Approach)*

Sources	Terms	Interpretations
Bubenko and Orci [1989]	Data	Assertions about perceived state-of-affairs in the UoD.
	Knowledge	Higher level of abstraction than data; concerned with properties, rules, and relationships in UoD.
	Information	Defined as the meaning of data.
Ramamoorthy and Wan [1989]	Data	Assertions about perceived states-of-affairs in the real-world.
	Knowledge	Meaning can be imprecise or uncertain since it may require an infinite amount of data to determine its truth.
Wiederhold [1986a and 1986b]	Data	Statements about specific instances (extensional).
	Knowledge	1) Statements about general concepts (intensional); 2) includes abstractions and generalizations of voluminous material.

2.3 Syntactics

The syntactics approach is concerned with how the syntax of a formal language is applied to represent a UoD. A syntax is a set of rules by which the signs of a language can be strung together. The language itself is defined as the collection of all the possible objects that can be constructed according to the syntax. Different models of a UoD can be constructed from the symbols of a language. The relevant issues in doing so include how to assemble symbols and take them apart, and what operations can be performed on a model. Note that these refer only to the syntactical aspect of a language, and ignore the real-world to which they refer.

2.3.1 Bubenko and Orci

Bubenko and Orci [1989]'s database view considers data to be a special case of knowledge. This is probably motivated by the current capabilities of typical DBMS's (for example, relational), together with the observation that relational databases are equivalent to a certain fragment of first-order logic, which we call DBFOL. (See Brachman and Levesque [1986].) They distinguish between data and knowledge in that data correspond to well-formed-formulae (wff) in DBFOL, whereas those statements requiring a more expressive language than DBFOL are knowledge. Consequently, databases represent sets of definite atomic statements (facts) and knowledge bases may, in addition, represent sets of general and conditional statements (rules).

This syntactical view is useful from an implementation standpoint since, if DBFOL suffices

for modelling the knowledge to be represented, then a DBMS can be used. Otherwise, some other approach is required.

2.3.2 Corella

Corella [1986] suggests that Wiederhold's distinction between data and knowledge [1986a] corresponds to the distinction between propositional logic (e.g., "Student Paul's age is 30 years") and quantified logic (e.g., "Every MBA student must take at least 60 credits to graduate"). Under this approach, data describes facts about an individual that can be represented by atomic formulae, whereas knowledge is a fact about a class of individuals and must be represented as a quantified well-formed-formula.

2.3.3 Forsyth

Forsyth [1989] focuses on representational characteristics of data and knowledge when examining the differences between a knowledge base and a conventional database. He identifies four salient properties for structures which represent data: (1) *Static*: the size of the structure is fixed. (2) *Flat*: the data contains atomic information which is not meant to be further subdivided; therefore, a piece of data may not contain substructures. (3) *Homogeneous*: all elements have the same data-type; for example, all occurrences of "name" are represented in the same way. (4) *Passive*: stores data, without computational attachment.

Structures for representing knowledge require the following essential properties. (1) *Flexible*: structures for representing knowledge must be extensible. (2) *Layered*: multiple layers are required to support inheritance. (3) *Heterogeneous*: knowledge can be represented in different ways. (4) *Active*: structures must allow for the representation of rules, methods, and so forth. Thus, this appears to be an ambitious attempt to define a set of complete properties for the structures to represent data and knowledge.

2.3.4 Ramamoorthy and Wan

Ramamoorthy and Wan [1989] suggest that "Information refers to bits that are stored in computer memories. These bits include data, software, and knowledge represented in data and software." Thus, information can be thought of as referring to any computer processable representation without regard for what it represents. This interpretation, therefore, is a syntactical one.

2.3.5 Summary

Table 3 provides interpretations of data and knowledge under the syntactics approach. Interpretations of data and knowledge based on the syntactic approach are basically logical. Data is represented by atomic formulae, whereas knowledge is represented by quantified well-formed-formulae. Information can be defined to be a "message" in the language of the logic which can be derived from the current stock of data and knowledge (set of current well-formed-formulae).

Table 3: **Definitions** *(Syntactics Approach)*

Sources	Terms	Interpretations
Bubenko and Orci [1989]	Data	Well-formed-formulae of DBFOL; special case of knowledge.
	Knowledge	Statements that cannot be represented in DBFOL.
Corella [1986]	Data	Facts represent by atomic formulae.
	Knowledge	Corresponds to quantified well-formed-formulae.
Forsyth [1989]	Data	Representation structures are static, flat, homogeneous, and passive.
	Knowledge	Representation structures are flexible, layered, heterogeneous, and active.
Ramamoorthy and Wan [1989]	Information	Bits that are stored in computer memories.

3 DISCUSSION

It is clear from the preceding discussion that there does not exist a consistent set of characteristics that describe data, knowledge, and information. Based upon the understanding that has been gained from applying the semiotics framework, we introduce five features that reflect the essential themes in the work reviewed above. These are: ***acquisition, processing, justification, real-world relationship***, and ***representation***. Note that the first three are based on pragmatics, the fourth, semantics, and the fifth, syntactics. These five features are first defined, and then applied as the basis for distinguishing amongst data, knowledge and information. In order to do so, ***statements*** will be judged as either ***simple*** or ***complex*** on each of these dimensions.

Acquisition: *Acquisition* simplicity is determined by the skill level required in order for the issuer of a statement to be considered credible in making it. A statement is simple with respect

to acquisition if a "normal" person with an "average" education would be credible in making it.

Processing: *Processing* simplicity concerns the degree to which a statement (or set of statements) is directly usable in making a decision or taking an action. If processing is required in order to make a statement usable, then it is complex on the processing feature. If a statement can be directly used, by a particular user, then, it is simple on the processing feature. For example, in an employee database, the stored statements are processing complex for retrieval of the average employee salary, but processing simple with respect to the salary of a specific employee.

Justification: *Justification* simplicity concerns a statement's relationship with other statements, and, in many cases, is closely connected with *acquisition*. The focus of justification is *why* a statement should be true rather than acquisition's focus on *how* a statement came to be known or believed. A statement is justification simple (for a given system or person), if its justification rests on simple grounds for example, ordinary sensors (eyes, thermometers, etc.)

Consider, for example, accounting statements in a firm's annual report. One can read from the balance sheet that the assets of the firm are $1,000,000, liabilities are $900,000, and equity is $100,000. From an acquisition standpoint, it is reasonable to say that one saw these numbers in an annual report in explaining how this information was acquired. Their justification, however, is based on the income and cash flow statements of the firm and a set of accounting principles; for example, equity is equal to assets less liabilities.

Real-World Relationship: *Real-World Relationship* simplicity pertains to issues of how many real-world objects are involved and how certain the statement is. Propositional statements such as "John's employee number is 77890" and "Mary is married to Robert" are simple in these aspects. Quantified statements, however, are related to the world in non-simple ways; for example, the statement "There exists at least one senior manager in each department." requires considerable effort to verify that it holds in the real-world.

Representation: *Representation* simplicity is expressed relative to a particular language. The view of Corella [1986] is typical and corresponds closely to this notion. In the database literature, for example, the relational model (and corresponding propositional logic (DBFOL);

see Brachman and Levesque [1986]) is typically taken as the standard for simplicity. Statements that can be expressed in the DBFOL are then simple; all others, by definition, are complex.

3.1 Data

We propose that statements (for example, facts or assertions) which are "simple" on each of the above five features (acquisition, processing, justification, real-world relationship and representation) should be called *data*. For example, facts that can be objectively verified (Wiederhold's interpretation of data) are, then, by definition, simple on acquisition, justification, and real-world relationship. They can be simply represented by, for example, attributes in a relational model. Processing of facts (to obtain other facts) would involve a simple operation such as "select" or "project".

3.2 Knowledge

Statements which are complex on at least one of the preceding features will be called knowledge. Based upon the review of previous research, there is some consensus that knowledge can be thought of as "data at a higher level of abstraction" (e.g. Ramamoorthy and Wan [1989], and others). Shaw [1984] defines an abstraction as a simplified description, or specification of a system that emphasizes some details or properties of an object while suppressing others. Well-known examples of knowledge, in this sense, include: derived knowledge, data abstractions, and induction, each of which is discussed briefly below with respect to our five features.

Derived Knowledge: Suppose that "Most employees in Department 10 are part-time" and that "Most part-time employees earn less than full-time". Then, a piece of derived knowledge, is "Most employees in Department 10 earn less than full-time employees". This statement is simple on the acquisition feature because most users would be able to draw this conclusion from the given premises. On the representation feature, however, it is complex because in the relational model (for example), there are not good mechanisms for capturing quantifiers such as "most". Note that this illustrates why uncertainty and judgment are often considered characteristics of knowledge. Thus, the statement is also complex on the real-world relationship and justification features. Finally, it is complex on the processing feature because more than a simple retrieval is required to respond to a query that would be interested in these results.

Data abstractions: A number of data abstractions have been identified by research in semantic data models. These are generally object-oriented and consist of the following types of primitive relationships amongst sets of entity types: inclusion (or *is-a*), aggregation (or *part-of*), and association (or *member-of*).

For example, the aggregation abstraction allows a relationship between objects to be thought of as a higher-level, *composite* object. This makes it possible to focus attention on the composite (or aggregate) while suppressing low-level detail [Smith and Smith, 1977]. This is clearly complex with respect to the real-world relationship feature because the relationships are on an entire class of entity types, and are, thus, quantified statements. Therefore, considerable effort will be required to verify that they hold true in the real world. Since at least one feature is complex, these abstractions are classified as knowledge.

Induction: Knowledge is sometimes obtained by making statements that serve to generalize a given body of data; for example, "all employees are over 21 years of age". This is done because it is desirable to make this piece of knowledge processing simple for whoever is interested in it. Such a piece of knowledge is drawn from a sample of the relevant population, only, and thus, is subject to uncertainty. Thus, it is complex on the relationship to the real world feature (because it requires observation of all the population in the real world to verify that it is true.)

3.3 Information

Information, similar to knowledge, has features that are complex. Information is "news", conveying material previously unknown to the receiver [Wiederhold, 1984; 1986b] and, therefore, in contrast to knowledge, requires a sender and a receiver (not necessarily distinct). Furthermore, a statement can only be news if it is understandable by the receiver [Orci and Bubenko, 1989]. Information can probably be best explained by examining why something is informative to the receiver. Suggestions for why this is so are as follows.

- Statements that are acquisition complex to the receiver are news because the receiver cannot carry out the original observation; for example, one obtains Consumer Reports because it is either difficult or impossible for an individual to obtain and/or compile the involved statements.
- A statement can be information because the original statements from which the information is derived are processing complex; for example, an investor's average rate of return

on a particular investment.

- With respect to justification, information can be simple or complex. For example, "800 employees will be hired this year due to the opening of a new plant", is justification simple. "All new employees have at least a Master's degree because only the R&D department does not have a hiring freeze" is justification complex.

- Information can be either simple or complex with respect to its relationship to the real world. For example, "Employee T. Willis is promoted to be Vice President" is simple; "All employees will receive a 10% wage increase this year" is complex (contains a universal quantifier).

- Information can be either simple or complex on the representation feature. For example, an employee's starting salary is information to the employee and can be represented using a simple numeric expression. "All employees will receive a 10% wage increase this year, with the exception of the Accounting Department" requires universal quantification, conjunction, and negation.

3.4 Relevance to System Development

The five features should be useful for developers of database and knowledge-based systems as explained below.

Acquisition. Do users need or want to know where the statements that are stored in a system come from? If the statements that a system stores are acquisition simple, then a user will probably not be concerned with their source. If, however, the statements are acquisition complex, then some explanation of the source may be important. This leads to two requirements determination activities. The analyst needs to: 1) understand the source of the system's stored statements; and 2) determine, from the users, whether knowing the source is important. In most existing systems, these issues are not usually treated in a systematic manner.

Processing. Here, the focus is the stored statements in a system versus users' queries. If users are interested in attributes of entire groups of things (e.g. statistics) as opposed to the underlying, individual statements, then there are two consequences for requirements determination. First, should the raw, underlying statements be stored? If so, then a set of processing

requirements must be determined in order to satisfy user queries. (Note, that the stored statements are then processing complex.) Second, instead of storing the underlying statements, one can store the summaries directly so that queries do not require significant, intermediate processing. In other words, there is some pre-processing on input. Therefore, the pre-processing requirements must be determined along with the appropriate structures for the summarized statements. In this case, the stored statements are processing simple with respect to user queries.

Justification. Current knowledge-based systems have mechanisms for explaining their reasoning process. However, neither database nor knowledge-based system are capable of justifying why the statements they actually store are true.

Real-World Relationship. The greater the real-world complexity of an application's stored statements, the greater the effort required on the part of the developer to ensure that the real-world is truly captured. This implies that: 1) statements stored in a system should be true in the real-world; and 2) a true real-world statement which is important to an application should be captured in a system.

Representation. This feature suggests that there is a need to have a common language in which the expressive power of alternative implementation schemes (e.g. relational model, object-oriented models, etc.) can be compared. Then, by expressing the statements that need to be stored for an application in this language, a developer can determine which implementation scheme is most appropriate for a given application.

4 SUMMARY & CONCLUSION

A number of different interpretations of data, knowledge, and information exist. A semiotics framework has been presented as a basis for analyzing these terms with respect to the current literature on databases and knowledge-based systems. The semiotics framework proved to be valuable in this analysis, because it highlighted five features as being most prominent. These features were defined in detail and then employed to distinguish amongst data, knowledge, and information. It is hoped that the results of this research will provide a clearer understanding of these terms, and, hence, how they should be used in the future development of database

management systems and knowledge-based systems.

References

[1] Bell, D., "The Social Framework of the Information Society", in Dertouzos, M.L. and Moses, J., (Eds.), *The Computer Age: A Twenty-Year View*, Cambridge, Massachusetts, The MIT Press, 1979.

[2] Brachman, R.J. and Levesque, H.J., "What Makes a Knowledge Base Knowledgeable? A View of Databases from the Knowledge Level", in Kerschberg, L. (Ed.), *Expert Database Systems*, The Benjamin/Cummings Publishing Company, 1986, pp. 69–78.

[3] Bubenko, J.A. and Orci, I.P., "Knowledge Base Management Systems: A Database View", in Schmidt, J.W. and Thanos, C. (Eds.), *Foundations of Knowledge Base Management*, Springer-Verlag, New York, 1989, pp. 373–378.

[4] Carnap, R., *Introduction to Semantics*, Cambridge: Harvard University Press, 1942.

[5] Corella. F., in Brodie, M.L., "Knowledge Base Management Systems: Discussions from the Working Group", in Kerschberg, L. (Ed.), *Expert Database Systems*, The Benjamin/Cummings Publishing Company, 1986, pp. 19–33.

[6] Forsyth, R., "From Data to Knowledge" in Forsyth, R. (Ed.), *Expert Systems: Principles and Case Studies*, Chapman and Hall Computing, New York, 1989, pp. 125–141.

[7] Hayes, P. J., "Some Problems and Non-Problems in Representation Theory", in Brachman, R.J. and Levesque, H.J. (Eds.), *Reading in Knowledge Representation*, Morgan Kaufmann Publishers, Inc. Los Altos, Califorina, 1985, pp. 3–22.

[8] Morris, C.W., *Signs, Language and Behavior*, Prentice-Hall, New York, 1946.

[9] Newell, A., "The Knowledge Level", *Artificial Intelligence*, 18:1, 1982, pp. 87–127.

[10] Ramamoorthy, C.V. and Wan, B.W., "Knowledge and Data Engineering", *IEEE Transactions on Knowledge and Data Engineering*, 1:1, 1989, pp. 9–16.

[11] Shannon C. and Weaver, W., *The Mathematical Theory of Communication*, the University of Illinois Press, 1962, reprinted from the Bell System Technical Journal, 1948.

[12] Shaw, M., "The Impact of Modelling and Abstraction Concerns on Modern Programming Languages", in Brodie, M.L., Mylopoulos, J., and Schmidt, J.W. (Eds.), *On Conceptual Modelling*, Sringer-Verlag, 1984, pp. 49–78.

[13] Smith, J. and Smith, D., "Database Abstractions: Aggregation and generalization", *ACM Transactions on Database Systems*, 2:2, June 1977, pp. 105–133.

[14] Stamper, R., *Information in Business and Administrative Systems*, C. Tinling & Co. Ltd., London, 1973.

[15] Wiederhold, G., "Knowledge and Database Management", *IEEE Software*, January 1984, pp. 63–73.

[16] Wiederhold, G., in Brodie et al., "Knowledge Base Management Systems: Discussions from the Working Group", in Kerschberg, L. (Ed.), *Expert Database Systems*, The Benjamin/Cummings Publishing Company, 1986a, pp. 19–33.

[17] Wiederhold, G., "Knowledge versus Data", in Brodie, M.L. and Mylopoulos, J. (Eds.), *On Knowledge Base Management System*, Springer-Verlag, New York, 1986b, pp. 77–82.